DACS-USC Data Repository: A Joint Effort by Quanterion and the University of Southern California (USC)

DACS Quanterion Team Member: Thomas McGibbon

DACS USC Team Members: Jo Ann Lane, Qi Li, and Ramin Moazeni

DRAFT as of 10 June 2010

Table of Contents

1. Introduction

The goal of the DACS-USC data repository is to capture and analyze software and software engineering data that will be used to improve:

- Quality of software–intensive systems
- Ability to predict the development of software–intensive systems with respect to effort and schedule.

This repository will:

1. Provide searchable data to support
    a. Cost estimation: rough order of magnitude estimates[1] based upon
        - Actual data – at least four records with no one organization providing more than 50% of the data points
        - Software size and application domain
    b. Project planning and management: life cycle model information, key risks, lessons learned, templates, estimation heuristics, software quality trends/data
2. Support software, systems, and system of systems (SoS) engineering research.

1.1 DACS-USC Contents

The repository will contain data primarily related to the development of software-intensive systems. This includes data related to the software, software engineering, systems engineering for software-intensive systems, and SoS engineering (SoSE) for net-centric, software-intensive SoS. Overall plans are for the inclusion of the following:

- **Software engineering, systems engineering, and SoSE** : **Size, schedule,** total cost of ownership, interoperability, and quality data/trends
- **Data entry forms**
- **Administrative data management tools**
- Lessons learned (acquirer and supplier perspectives)
- Counting rules for quantitative data
- Software/system characteristics: Life cycle models, architecture styles, tools used, tool evaluations
- Process-related artifacts: Templates, characteristics, home-grounds

The bolded items are those that are planned for the initial version of the repository.

[1] Repository will provide links to participating cost model vendors for more detailed/precise estimates and additional support.

1.2 Potential DACS-USC Data Sources

Figure 1 identifies candidate data sources for the repository.

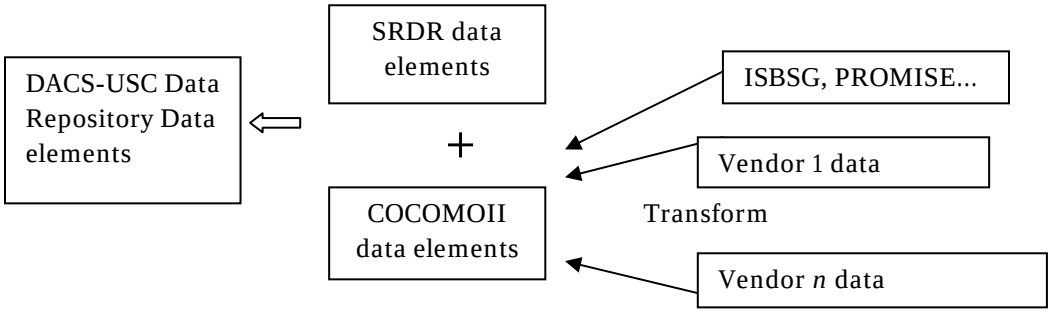


Figure 1

The two sources planned for the initial version of the repository are sanitized COCOMO™ projects that are approved for incorporation through USC CSSE agreements with the data owners and data from the Department of Defense (DoD) Software Resources Data Report (SRDR) submissions from software system developers. Other generally available data such as ISBSG and PROMISE are currently being evaluated for inclusion into the repository. In addition, inputs are being solicited from cost model vendors as well as other organizations that have actual data from completed projects. Actual data in repository will depend on ability to find sources for the various data elements.

In general, all data will be stripped of any information that might indicate the original data source, project, or developer and any necessary identifying information needed to maintain the contents of the repository will be kept offline in protected media. (Note: Repository data is only displayed in an aggregated form where there are at least four projects in the aggregation and no one organization has over 50% of the data sets. This is in accordance with the USC Center for Systems and Software Engineering Data Management and Security Procedures provided in Appendix A.)

1.3 DACS-USC Data Repository Architecture

The architecture of the DACS data repository will be a 3-tier architecture comprised of open source products to the extent possible. The general structure of the 3 tiers is:

Client Tier: HTML+CSS

Middle Tier: PHP

Backend Tier: MySQL

More detailed information on the repository architecture and security strategies is available in the DACS-USC Data Repository Project Plan that has limited distribution at this time.

2 Overview of Currently Prototyped Functionalities

The current DACS-USC DATA REPOSITORY prototype focuses on the Software Data, including functionalities "Searching for similar software projects and show statistics report" , "Import software data" and "display detailed project information" (only for administrator). Capabilities related to system engineering and SoSE data will be implemented in later iterations.

2.1 Data Search and Report

After the user logs into the DACS-USC Data Repository (repository) homepage as shown in Figure 2, the left bar provides the instructions on the above "*Dear**, Welcome to DACS-USC Repository! Choose Software, System or System of System Data below to find similar projects in this repository, if you want to contribute your organization's data, please go to Contribute Data Section.* " and lists the primary repository feature links. Click on the "Software Data" below "Data Repository" and it will lead to the software data query page in Figure 3.

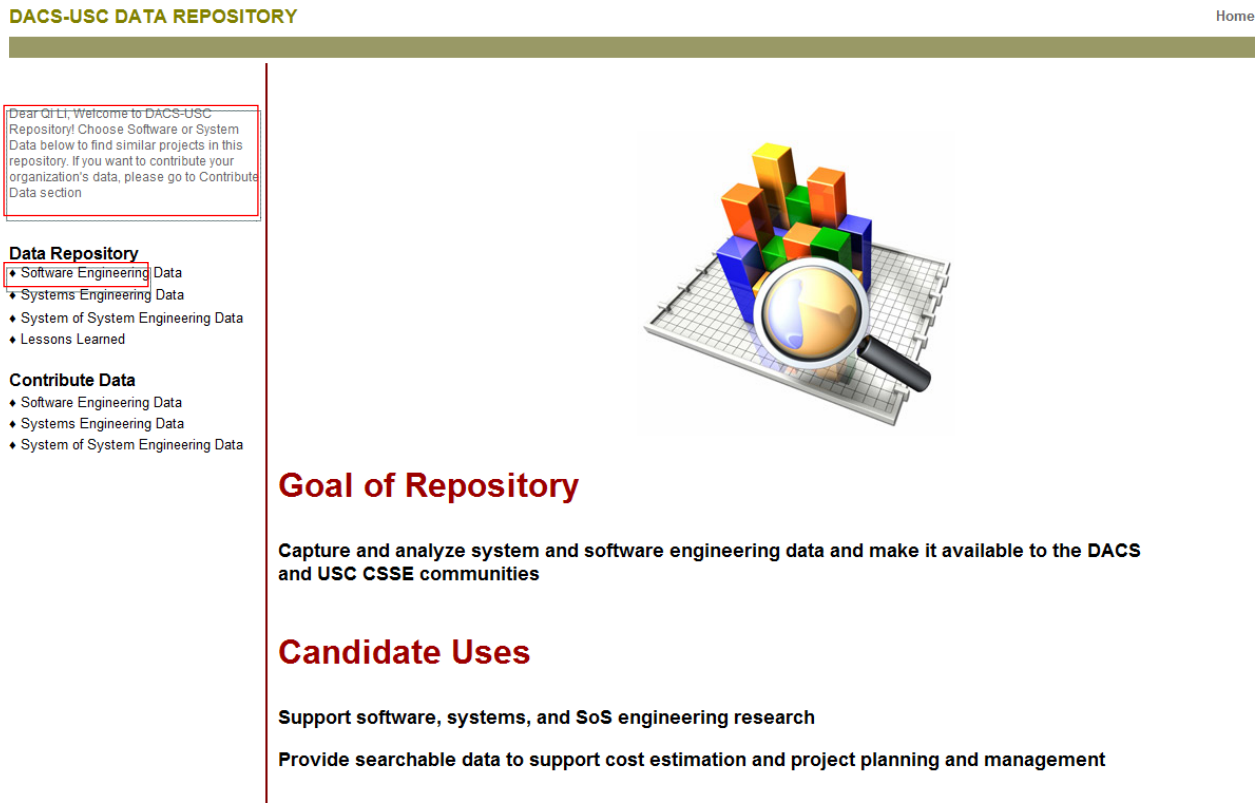


Figure 2

In the software data query page in Figure 3, users can currently query existing software data by "Size" and "Domain" as displayed on top of this page.

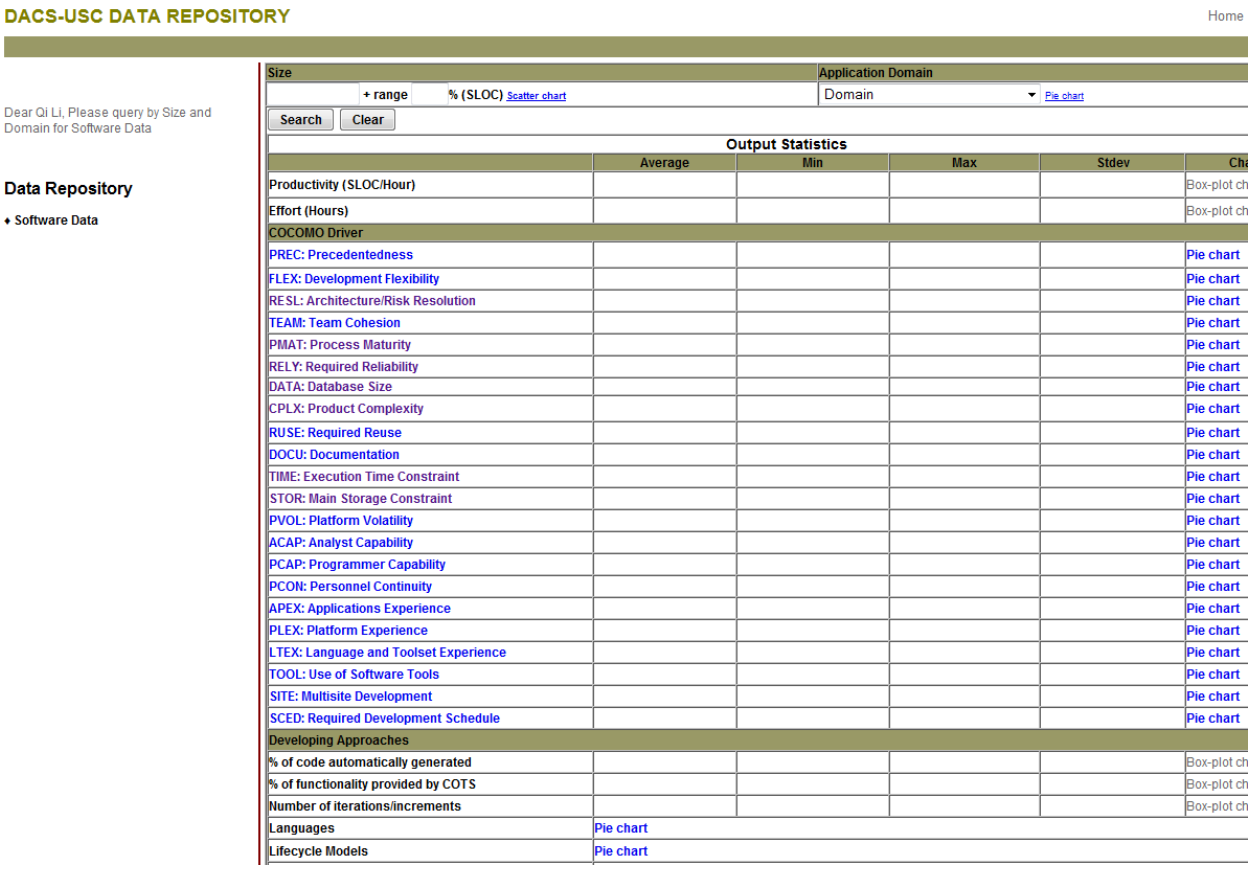


Figure 3

After the user inputs the query conditions, and clicks on "Search", the statistics results (Average, Min, Max and Standard Deviation) and related charts for all the measures (Productivity, Effort, Schedule, COCOMO drivers and Developing Approaches) are displayed in the table below. If you would like to see the definition of the measure or how it was calculated, you can click on it, and it will provide you a detailed explanation as shown in Figure 4. A repository search example is displayed in Figure 5 to Figure 9.

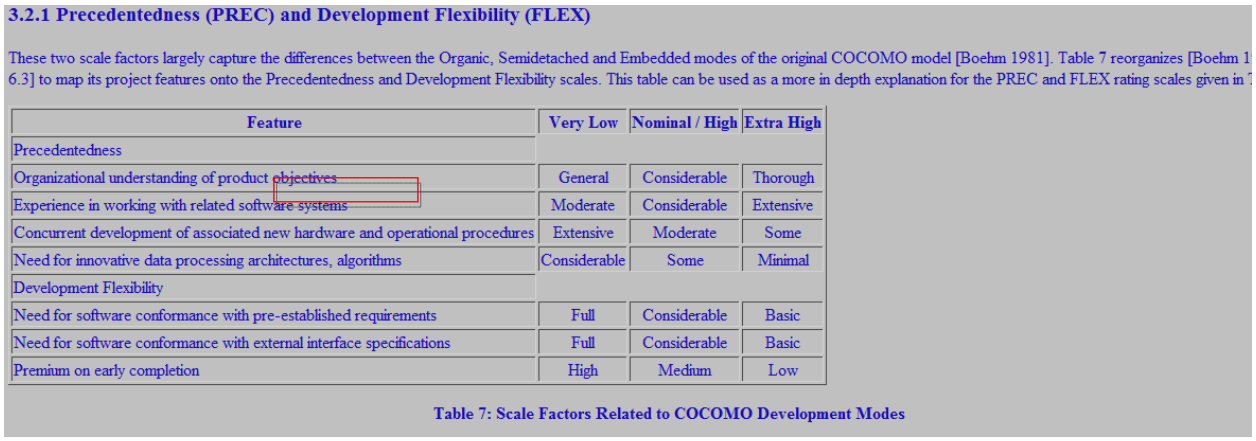
Topics for further discussion:

- *What other query conditions are needed in order to find similar projects in the repository?*
- *What other measures do we need to show? (This also means what other data attributes the data contributor should provide? Which are required, which are optional?).*
- *What other statistics or charts do we need to show the searched result?*

Other candidate search parameters for similar projects within the repository:

1) *Size with ranges*
2) *Domain*
3) *Main language used*
4) *Development type: new or update*

5) *Cost driver values:*
   a. *RELY*
   b. *CPLX*
   c. *SCED*
   d. *Etc…*



3.2.1 Precedentedness (PREC) and Development Flexibility (FLEX)

These two scale factors largely capture the differences between the Organic, Semidetached and Embedded modes of the original COCOMO model [Boehm 1981]. Table 7 reorganizes [Boehm 1
6.3] to map its project features onto the Precedentedness and Development Flexibility scales. This table can be used as a more in depth explanation for the PREC and FLEX rating scales given in

Feature	Very Low	Nominal / High	Extra High
Precedentedness			
Organizational understanding of product objectives	General	Considerable	Thorough
Experience in working with related software systems	Moderate	Considerable	Extensive
Concurrent development of associated new hardware and operational procedures	Extensive	Moderate	Some
Need for innovative data processing architectures, algorithms	Considerable	Some	Minimal
Development Flexibility			
Need for software conformance with pre-established requirements	Full	Considerable	Basic
Need for software conformance with external interface specifications	Full	Considerable	Basic
Premium on early completion	High	Medium	Low

Table 7: Scale Factors Related to COCOMO Development Modes

Figure 4

A searching example:

Before the user searches for information about similar projects in the repository, he can first review general information about the repository contents by clicking "Scatter chart" (beside the Size input) as shown in Figure 5 and "Pie Chart" (beside the Domain input) as shown in Figure 6. In this example, 213 software data records are currently in the repository. The "Overview SLOC Effort Scatter Chart" provides an overview of the 213 projects' size and effort range in the repository. The "Overview Domain Distribution" Chart provides an overview of the projects' domain distribution in the repository. As shown in Figure 6, the current software database covers 13 domains and most software projects in the repository belong to the "Business" domain. The Business domain has a total of 45 records, consisting of 21.1% of all software projects.

Note: We should provide an authoritative list of domain categories. One solution is referring to SEER-SEM's domain list, other information is USC "AFCAA Database & Metrics Manual" team is now doing some research on this and we can communicate with them about this at ARR.

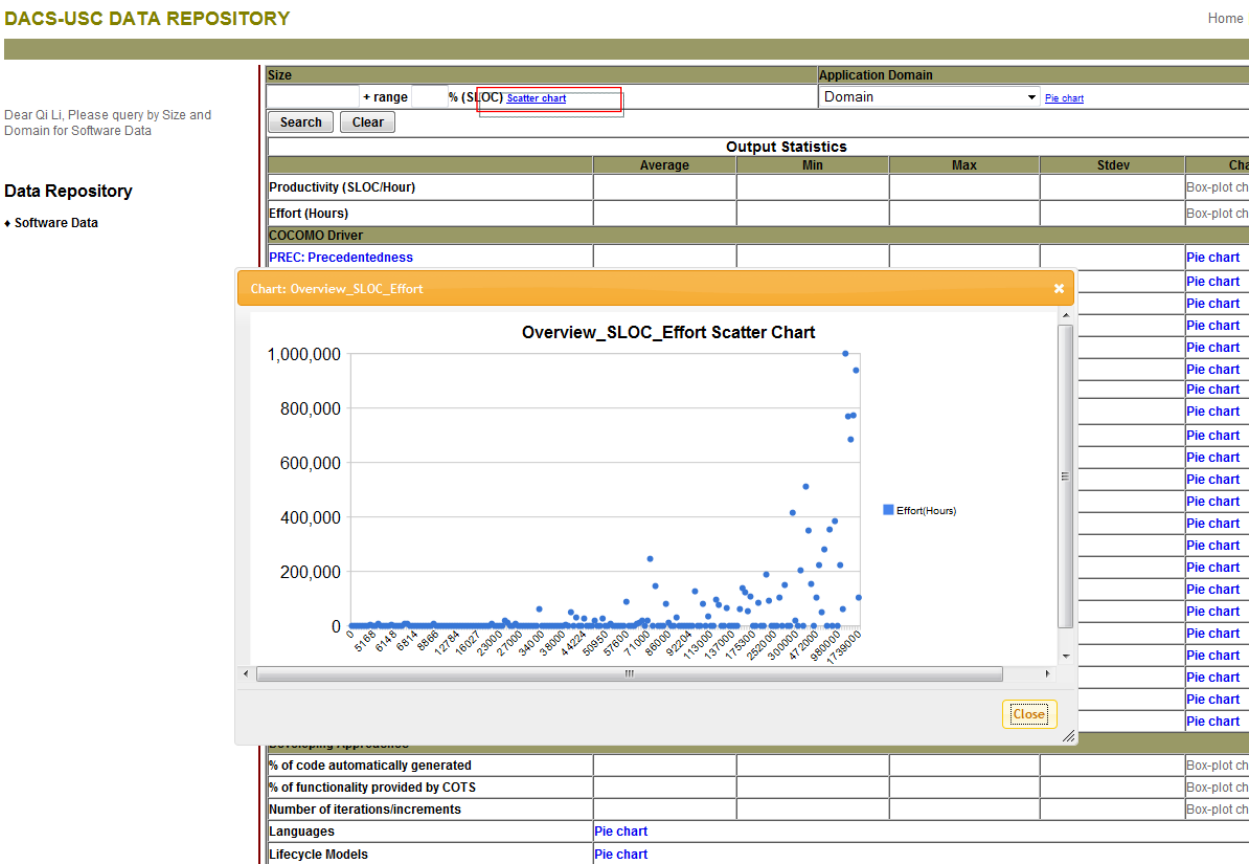

DACS-USC DATA REPOSITORY
Home
Dear Qi Li, Please query by Size and Domain for Software Data
Data Repository
♦ Software Data
Size
+ range
% (SLOC) Scatter chart
Application Domain
Domain
Pie chart
Search
Clear
Output Statistics
Average
Min
Max
Stdev
Productivity (SLOC/Hour)
Effort (Hours)
COCOMO Driver
PREC: Precedentedness
Pie chart
Box-plot ch
Chart: Overview_SLOC_Effort
Overview_SLOC_Effort Scatter Chart
1,000,000
800,000
600,000
400,000
200,000
0
Effort(Hours)
Close
% of code automatically generated
% of functionality provided by COTS
Number of iterations/increments
Languages
Lifecycle Models

Figure 5

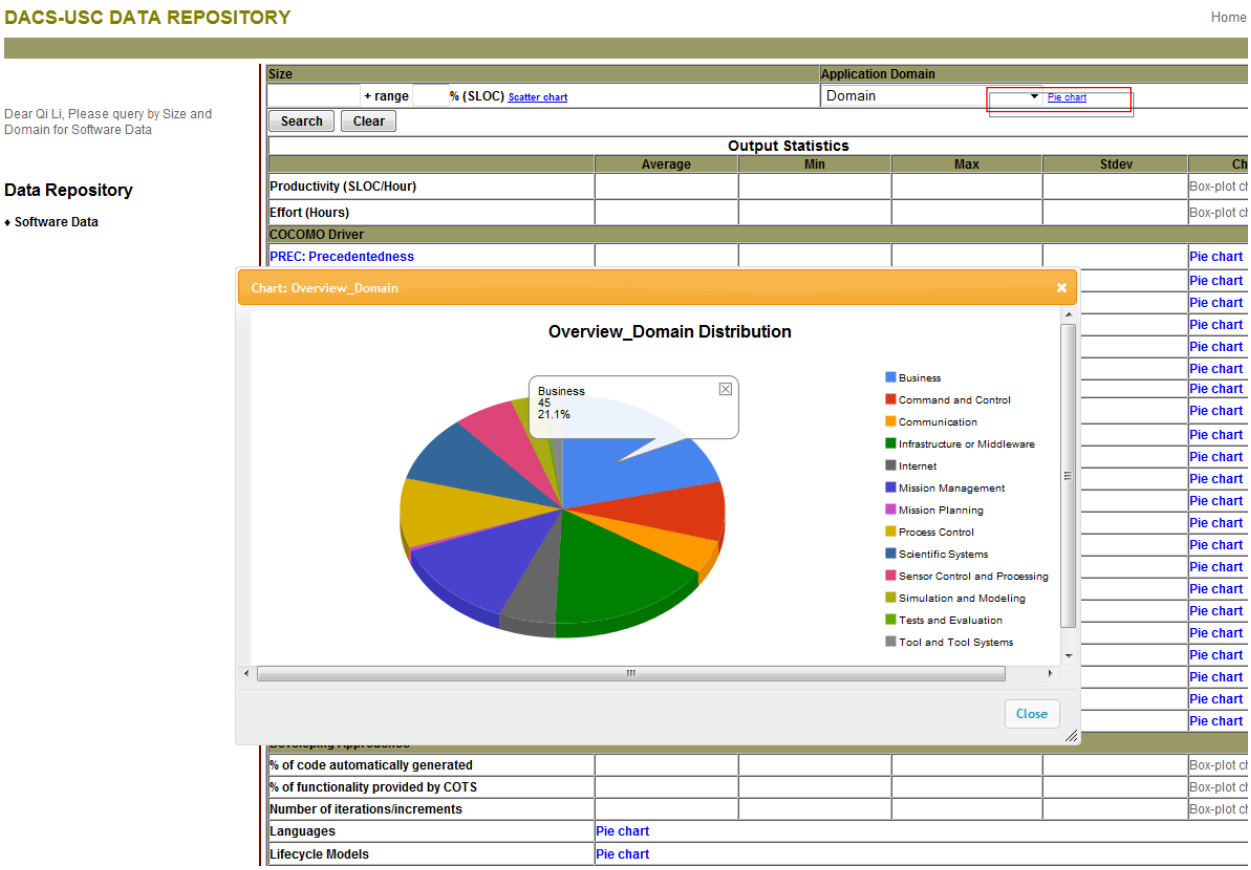


Figure 6

As shown in Figure 7, suppose that the user estimates his current project's size to be "60000" SLOC, but feels that the estimates have a lot of uncertainty. One option is for the user to widen the search range by 50%, meaning that the system should search for size in the range of 60000*(1-50%) to 60000*(1+50%) or [30000, 90000] SLOC. If the user's project belongs to the "Business" domain, the user would choose "Business" in the dropdown list.

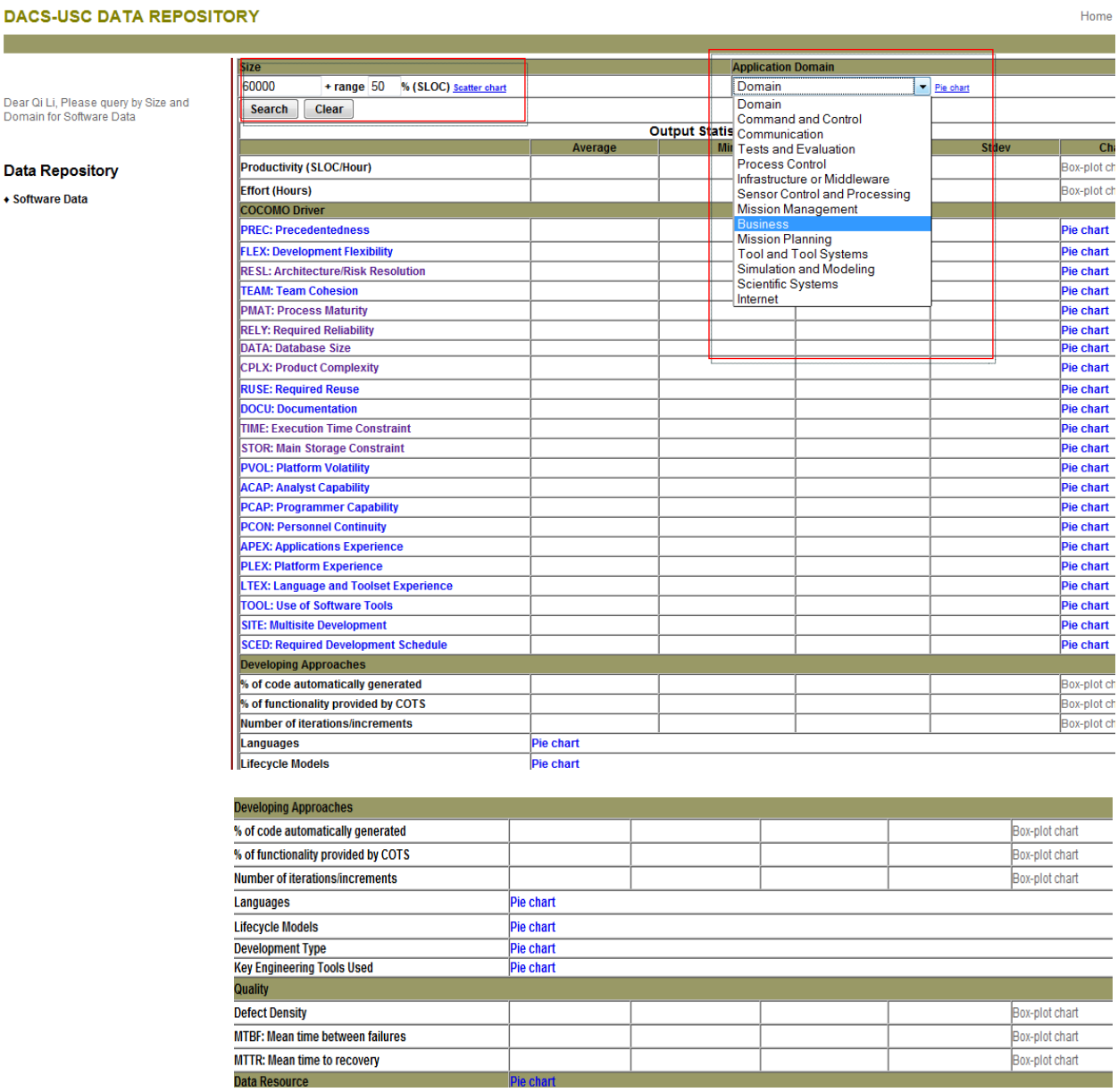
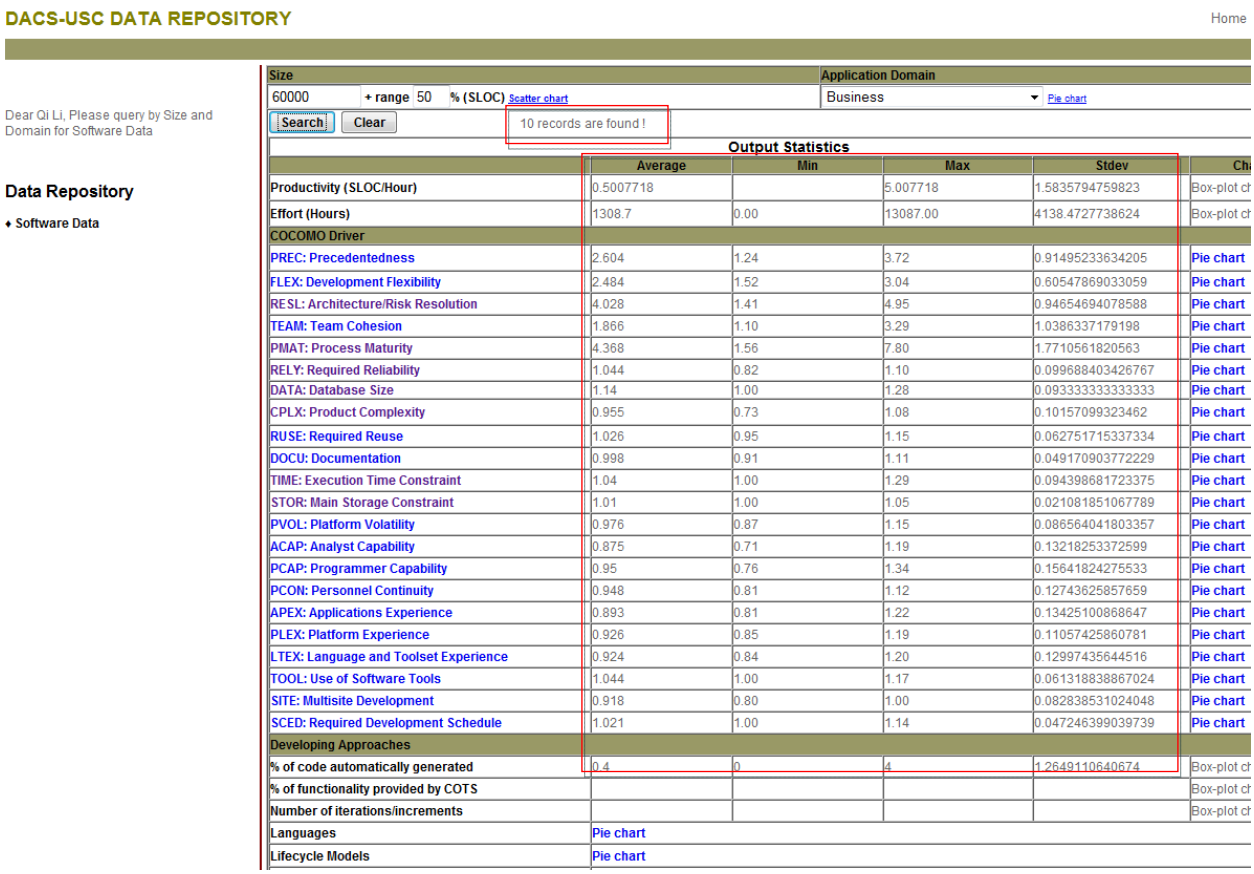


Figure 7

When the user clicks on "Search", it shows that "10 records are found!" and all the related statistics are displayed on the "Output Statistics" table, as shown in Figure 8. This information provides the user with some useful insights about the general productivity, cost drivers range, and development approaches used by other similar projects. This information can be used to support the user's project planning and estimating activities.



DACS-USC DATA REPOSITORY

Home

Dear Qi Li, Please query by Size and Domain for Software Data

Data Repository

- Software Data

Size: 60000 + range 50 % (SLOC) Scatter chart

Application Domain: Business Pie chart

Search | Clear

10 records are found !

Output Statistics

	Average	Min	Max	Stdev	Ch
Productivity (SLOC/Hour)	0.5007718		5.007718	1.5835794759823	Box-plot ch
Effort (Hours)	1308.7	0.00	13087.00	4138.4727738624	Box-plot ch
COCOMO Driver					
PREC: Precedentedness	2.604	1.24	3.72	0.91495233634205	Pie chart
FLEX: Development Flexibility	2.484	1.52	3.04	0.60547869033059	Pie chart
RESL: Architecture/Risk Resolution	4.028	1.41	4.95	0.94654694078588	Pie chart
TEAM: Team Cohesion	1.866	1.10	3.29	1.0386337179198	Pie chart
PMAT: Process Maturity	4.368	1.56	7.80	1.7710561820563	Pie chart
RELY: Required Reliability	1.044	0.82	1.10	0.099688403426767	Pie chart
DATA: Database Size	1.14	1.00	1.28	0.093333333333333	Pie chart
CPLX: Product Complexity	0.955	0.73	1.08	0.10157099323462	Pie chart
RUSE: Required Reuse	1.026	0.95	1.15	0.062751715337334	Pie chart
DOCU: Documentation	0.998	0.91	1.11	0.049170903772229	Pie chart
TIME: Execution Time Constraint	1.04	1.00	1.29	0.094398681723375	Pie chart
STOR: Main Storage Constraint	1.01	1.00	1.05	0.021081851067789	Pie chart
PVOL: Platform Volatility	0.976	0.87	1.15	0.086564041803357	Pie chart
ACAP: Analyst Capability	0.875	0.71	1.19	0.13218253372599	Pie chart
PCAP: Programmer Capability	0.95	0.76	1.34	0.15641824275533	Pie chart
PCON: Personnel Continuity	0.948	0.81	1.12	0.12743625857659	Pie chart
APEX: Applications Experience	0.893	0.81	1.22	0.13425100868647	Pie chart
PLEX: Platform Experience	0.926	0.85	1.19	0.11057425860781	Pie chart
LTEX: Language and Toolset Experience	0.924	0.84	1.20	0.12997435644516	Pie chart
TOOL: Use of Software Tools	1.044	1.00	1.17	0.061318838867024	Pie chart
SITE: Multisite Development	0.918	0.80	1.00	0.082838531024048	Pie chart
SCED: Required Development Schedule	1.021	1.00	1.14	0.047246399039739	Pie chart
Developing Approaches					
% of code automatically generated	0.4	0	4	1.2649110640674	Box-plot ch
% of functionality provided by COTS					Box-plot ch
Number of iterations/increments					Box-plot ch
Languages	Pie chart				
Lifecycle Models	Pie chart				

Figure 8

To show the statistical results graphically, charts are provided by the system. For example, Figure 9 shows the RESL rating distribution for the 10 projects identified by the query. It shows that 8 (80%) projects have an "Nh" rating for RESL. Later on in the project, a Box-plot chart may be implemented to show the statistics for productivity, effort etc.

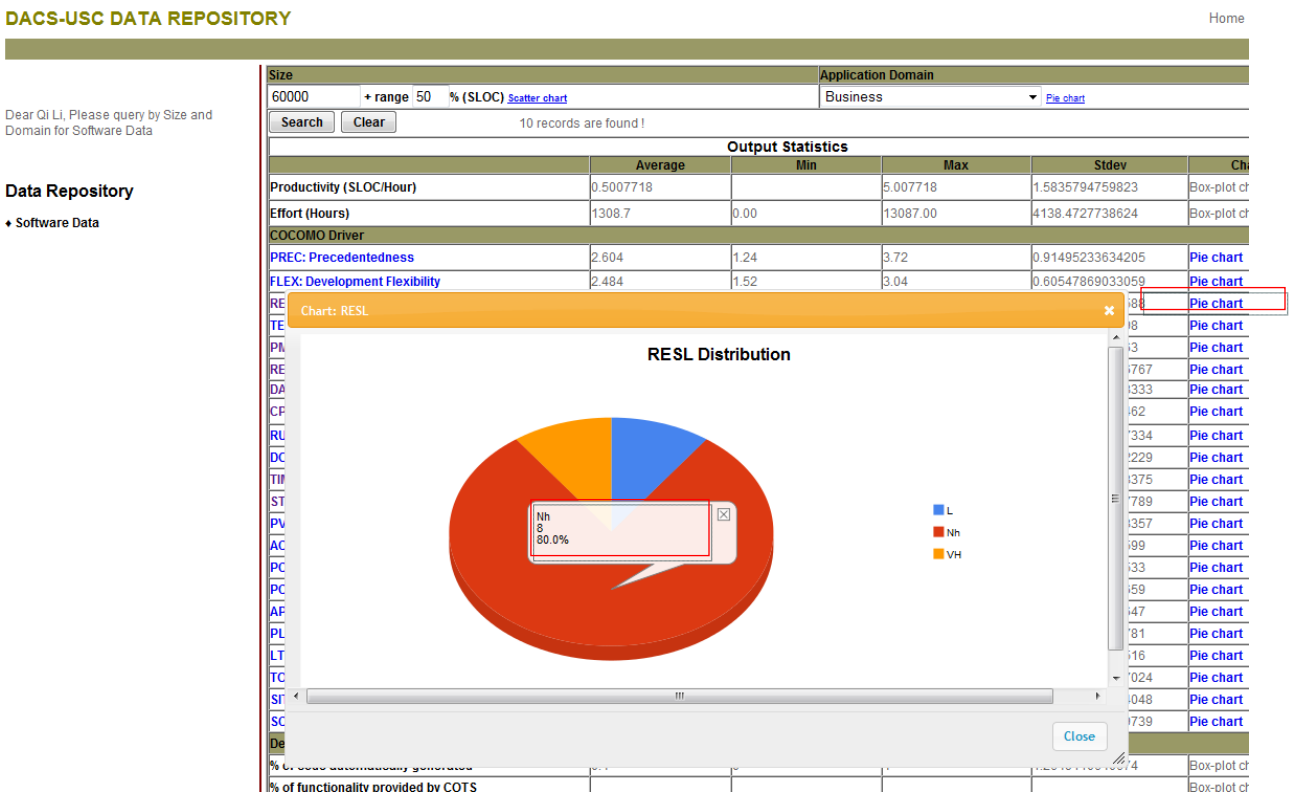


Figure 9

2.2 Report Formats (TBD)

This section will describe standard reports that can be requested from the repository. The requested report will be provided in a file that the user can save and print. Exact contents, formats, and file type currently TBD.

2.3 Lessons Learned (TBD)

This section will describe the types of lessons learned that will be provided in the repository as well as how the lessons learned will be categorized and tagged.

Current category considerations:

1. *Since we can use risks to capture lessons learned, we can also use risk categories for classifying lessons learned. The cost-driver factors in software cost-estimation models provide a checklist of risk items that are strongly correlated with cost and schedule increases on past projects. Please refer to Green Book (Rick Selby edited book) p429-436.*
    a. *Aspect of Personnel*
    b. *Aspect of Schedule & Budgets*
    c. *Aspect of Requirements*
    d. *Aspect of External Components and Tasks*
    e. *Aspect of Performance*
    f. *Aspect of Computer Science Capabilities*

2. *We can also refer to "The Macro Risk Model: An Early Warning Tool for Software-Intensive Systems Projects" (http://csse.usc.edu/csse/event/2009/UARC/material/Macro Risk Model.pdf) and add the following categories:*
    a. *Aspect of Process Management*
    b. *Aspect of Strategies*
3. *Furthermore, we can consider categories identified in "A Software Engineering Lessons Learning Repository" (by Warren Harrison) (http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C01D23197A34A5088F4DDA90DC74E37E?doi=10.1.1.11.1761&rep=rep1&type=pdf):*
    a. *Aspect of Domain*
    b. *Aspect of Phase (Requirement->Design->Code->Test->Transition)*
    c. *Aspect of Product and Functional Category*
4. *Other categories to consider:*
    a. *Aspect of Risk Management*
    b. *Aspect of Quality Management, including Review, Test and V&V etc.*
    c. *Aspect of Feasibility Analysis, e.g. Cost-Benefit analysis, ROI analysis, etc.*
    d. *Aspect of Project Planning*
    e. *Aspect of Effort/Cost/Resources Estimation*

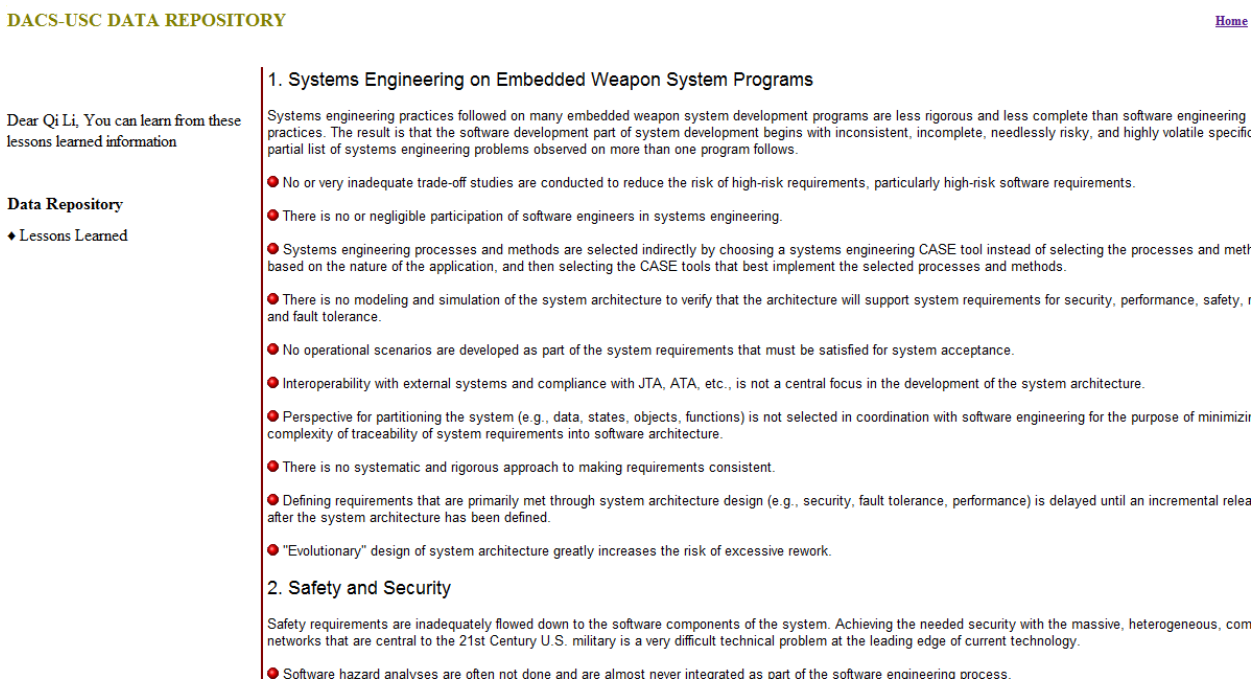


Figure 10 Example of potential lessons learned.

2.4 Data Submission (still under development)

The data submission user input screens are primarily based upon information available from the DoD SRDR forms. Note that in many cases, we have annotated the section of the SRDR form that the indicated data comes from.

Note: SRDR data elements are attached at the end of this document. More information is "AFCAA Database & Metrics Manual" (http://csse.usc.edu/afcaa/manual_draft/)---especially DATA ITEM DESCRIPTION.

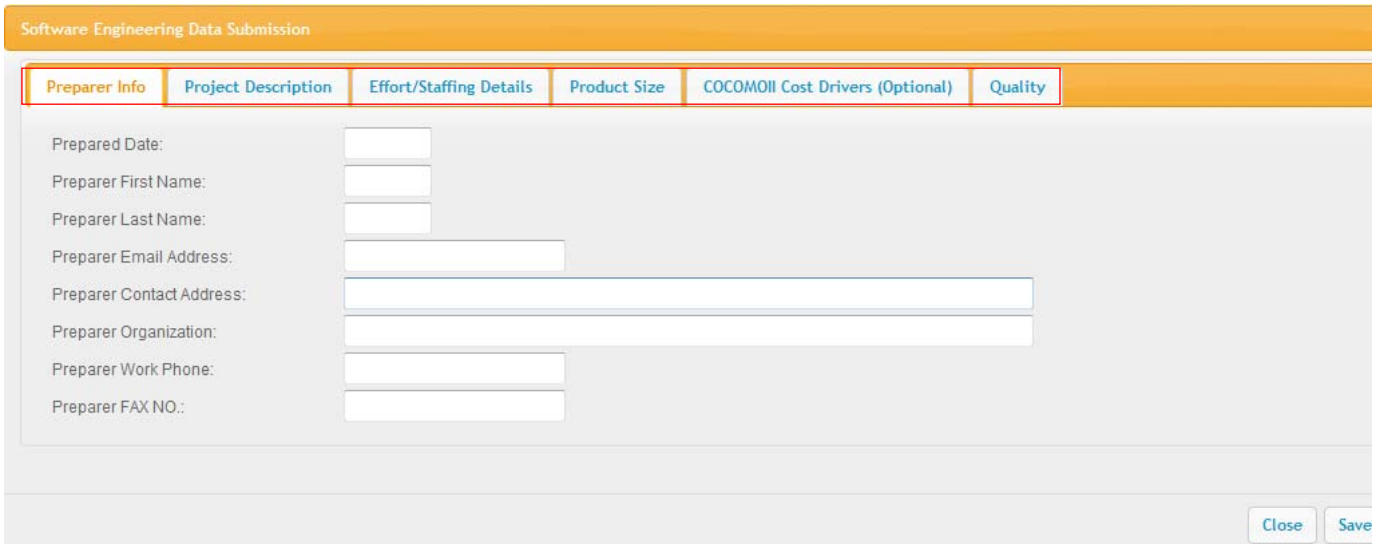


Figure 11 data elements categories

2.4.1 Preparer information:

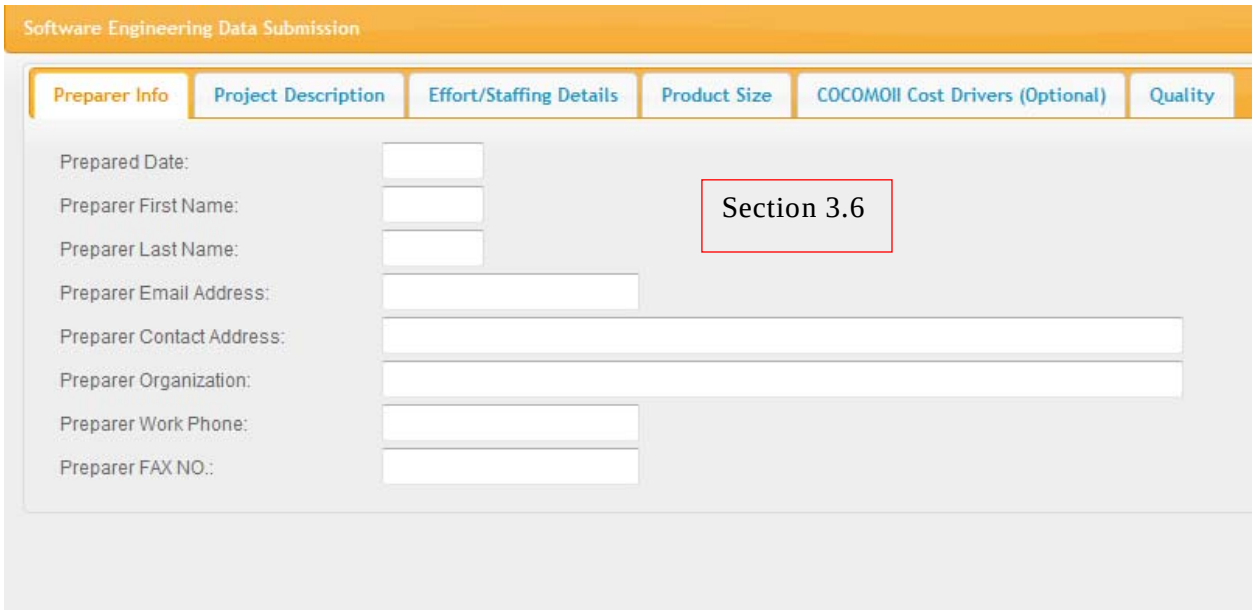


Figure 12 preparer information

2.4.2 Project description

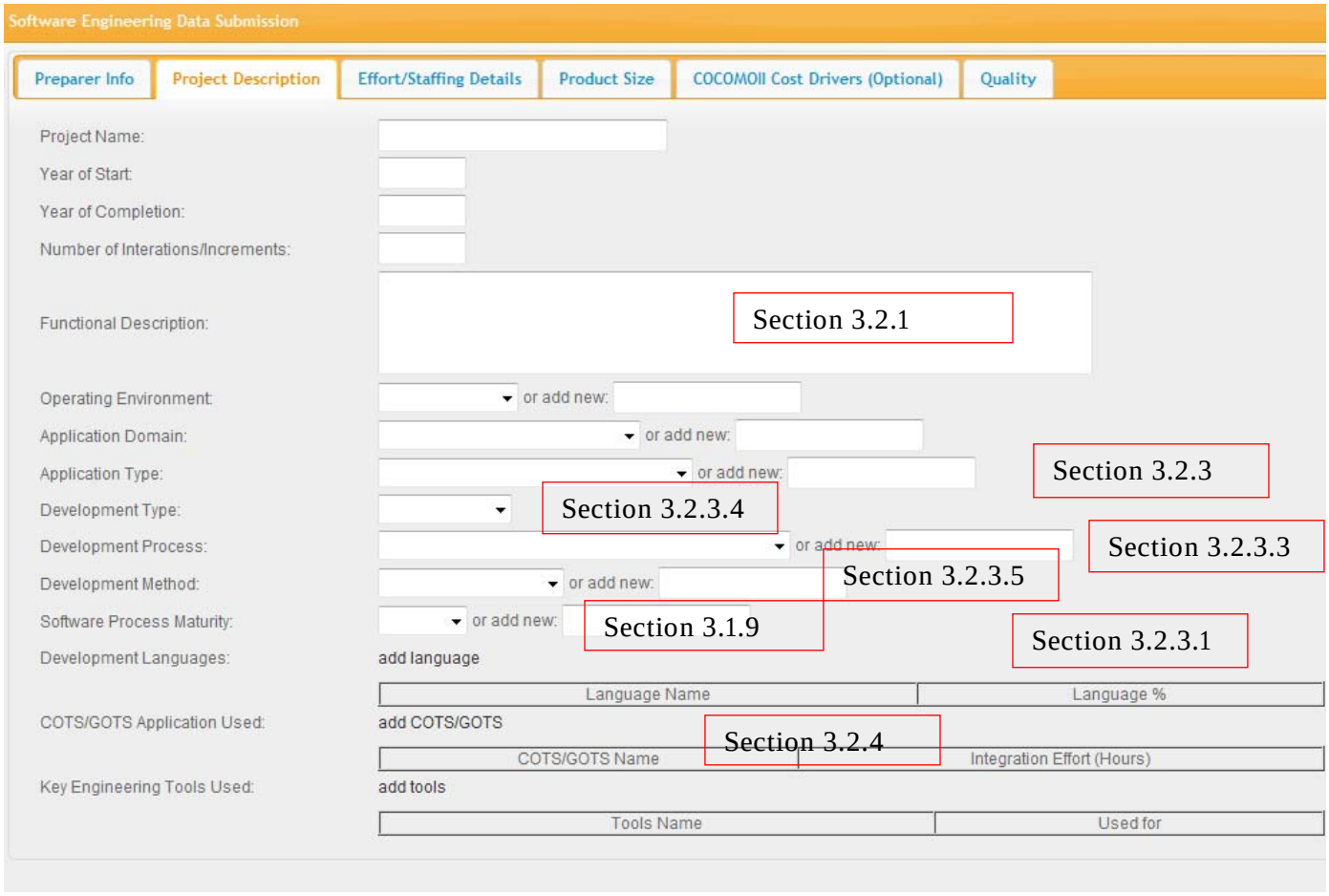


Figure 13 project description

2.4.3 Effort/staffing details

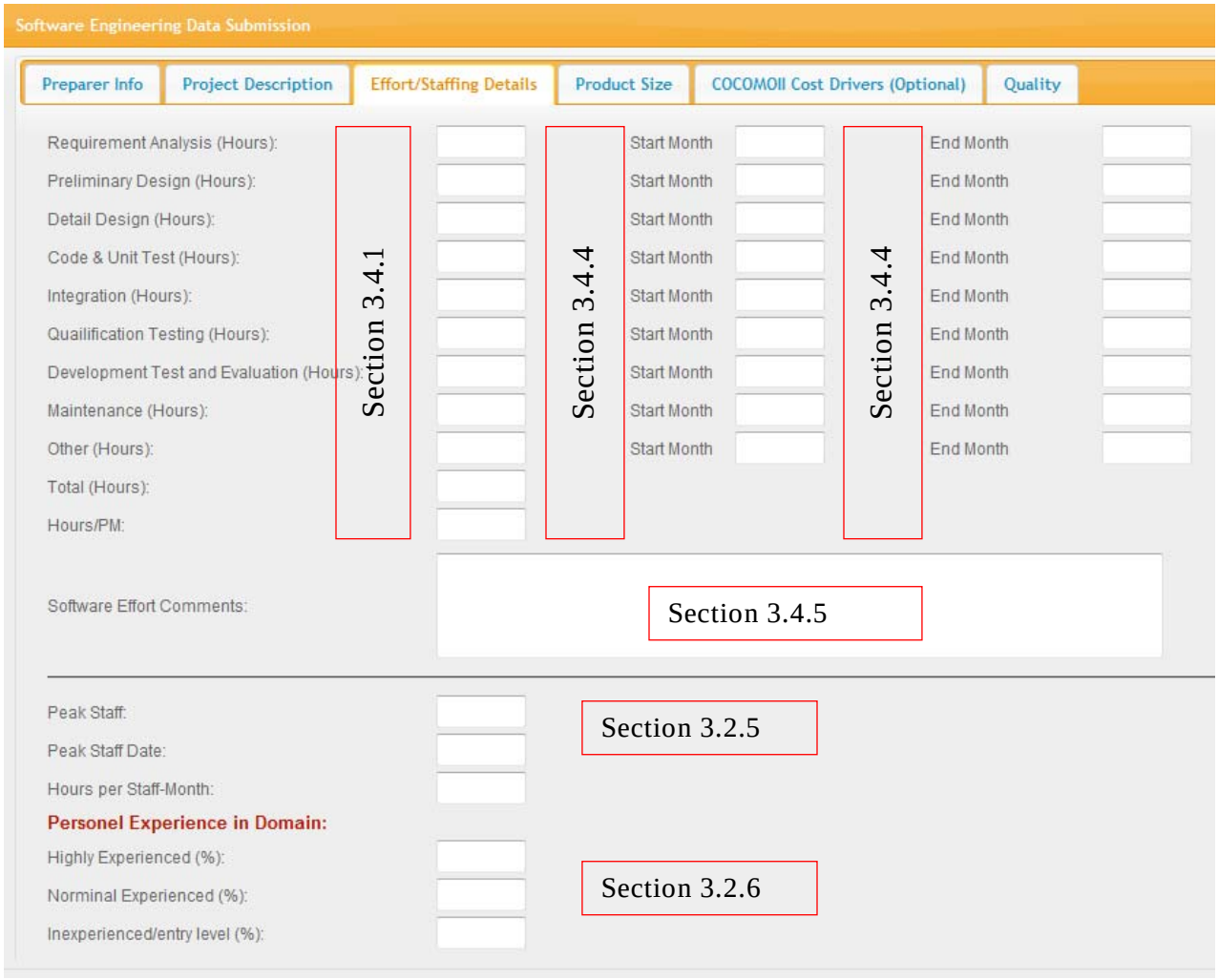


Figure 14 effort/staffing details

2.4.4 Product size

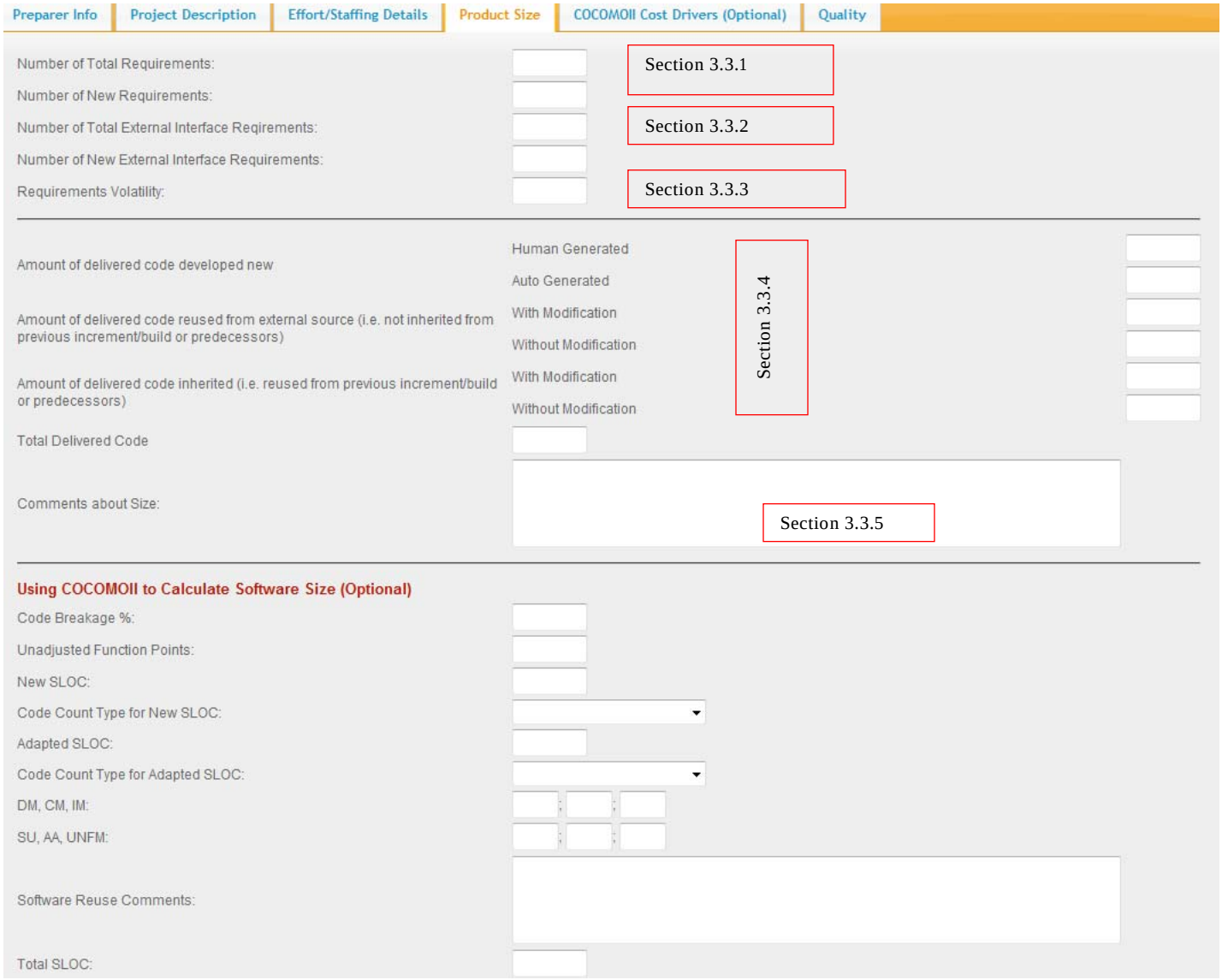


Figure 15 product size

2.4.5 COCOMOII Cost Drivers (Optional)

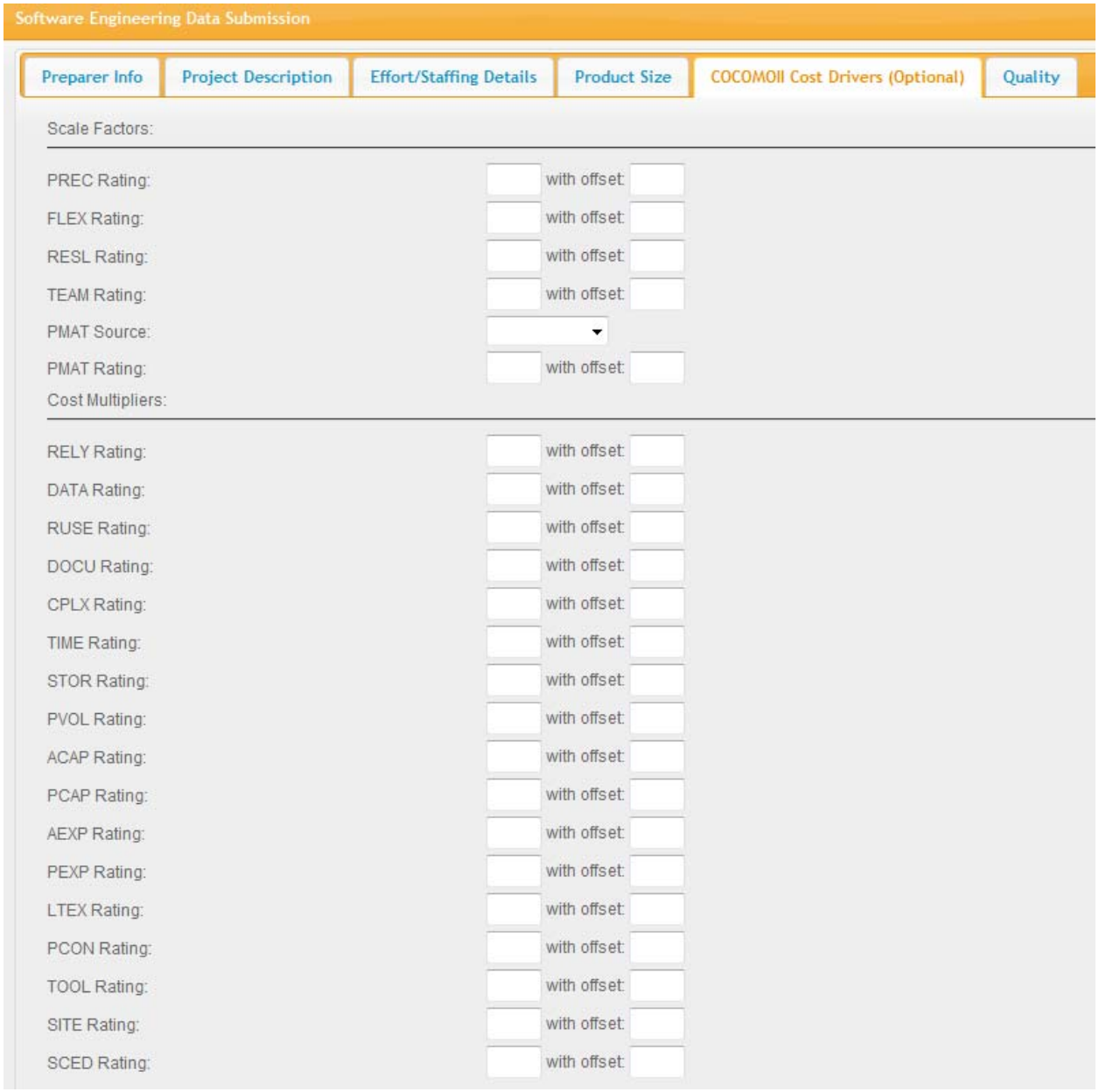


Figure 16 COCOMOII Cost Drivers (Optional)

2.4.6 Quality

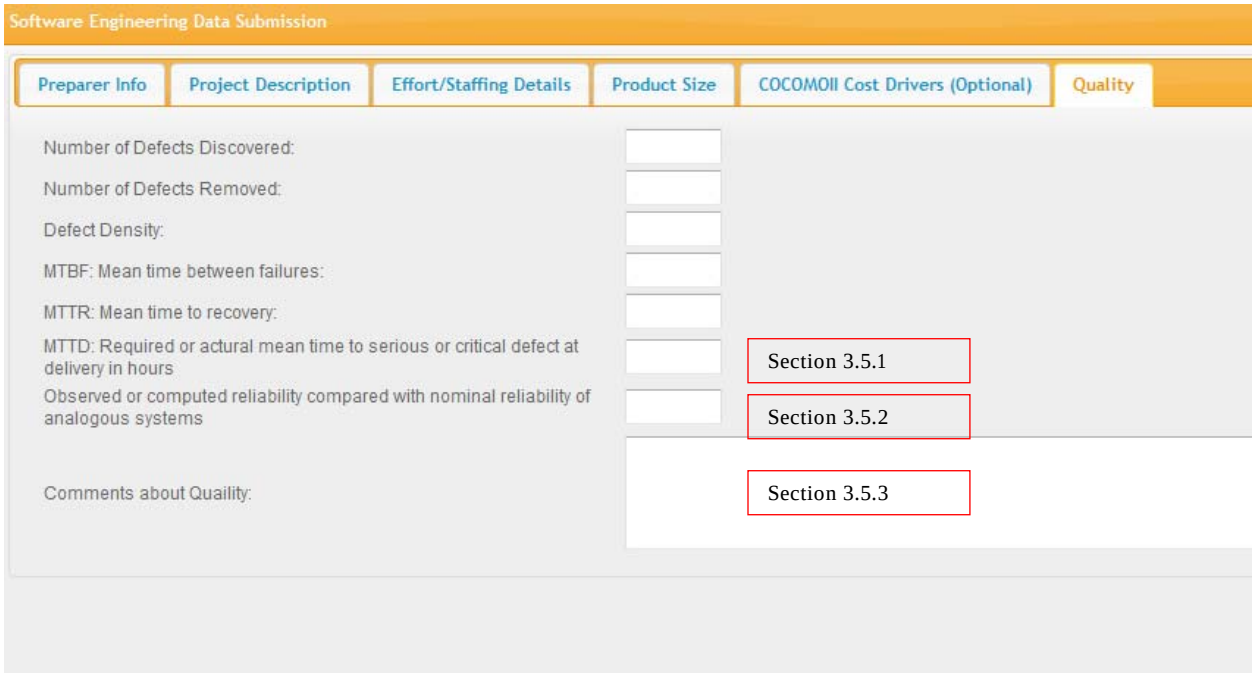


Figure 17 Quality Data from SRDR Form.

2.4.7 User Management System (TBD)

Will address various roles and relevant privileges (administrator, general users, do we have other roles, such as guest etc…)

Appendix A: USC Center for Systems and Software Engineering Data Management and Security Procedures

USC Center for Systems and Software Engineering Data Management and Security Procedures

This document details the procedures we, at the USC-Center for Systems and Software Engineering (CSSE), take to manage the software/systems project-related data we receive from our affiliates and clients. This agreement applies to all USC-CSSE data collection projects where the data may be considered sensitive or proprietary by affiliate or client data providers/submitters. It currently covers the COCOMO™ model suite of data and the DACS-USC Repository data. The procedures are covered as follows:

1. Data Identification
2. Data Submission
3. Data Storage
4. Data Access
5. Getting Help

1. Data Identification

Each Affiliate software project contributing data has a separate file identification number of the form XXX-YYY-NN-yyyymmdd.

XXX is one of a random set of three-digit organization identification numbers provided by USC to the Affiliates and only known to the Affiliate and Barry Boehm or Jo Ann Lane at USC CSSE. Note that a given Affiliate organization may have more than one identifier. For example, an Affiliate may request multiple IDs to distinguish various organizational sites or internal organizations.

YYY is a three-character alpha-numeric ID number assigned by the Affiliate when establishing a new project data file. Only the submitting Affiliate knows the correspondence between YYY and the actual project name.

NN is a two-digit number to distinguish multiple data submissions from a single project corresponding to different development cycles or iterations.

yyyymmdd identifies the date the data was submitted to the repository.

2. Data Submission

Data submission is an important aspect of this procedure. There is a need for data submission with the protection to the Affiliates' privacy. There is also a need to be flexible in delivery medium. In addition, the procedure used must be possible to detect unusual or erratic data and trace it back to its source.

2.1 Protection Level and Duration

Each submitter (identified by XXX) is asked to provide a "sunset" time at which point the data is no longer requires strong protection and sanitized versions (i.e., source/project identifying information not included) may be made available to others (e.g., the research community and USC CSSE clients and associates). Submitters can indicate a number of years or "none". Submitters should specify "none" for data they do not ever want made available to anyone outside of those on the USC CSSE data access list. If a "sunset" time is not provided by the submitter or other representative from his/her organization, it will be assumed to be 10 years from the date of submission. If a submitting organization gets acquired by another organization, the data will be protected in accordance with the acquiring organization's policy or for 10 years if there is no acquiring organization policy on file.

2.2 Format

The USC CSSE data collection projects would like the data submitted by using the Data Collection form provided or as COCOMO™ data files. Actuals can be captured in a spreadsheet or in a calibration file that is part of the COCOMO™ package. Alternatively, data can be submitted in other formats given that the format has been reviewed and approved by a member of the data access team.

3. Data Storage

The USC CSSE data collection project has obtained a secured room with limited access at USC for its data facility. The door is outfitted with a cypher lock on the door and is locked at all times. Only those personnel on the authorized list may enter the room and access the data. Each project's sanitized data[2] is maintained on at least two independent, portable hard drives that are stored in a locked filing cabinet in the secure room when not in use. At least two copies of the data allows for use by multiple authorized personnel as well as providing a reliable backup of the data.

Data submissions received via Internet electronic mail will be inaccessible once it is received and stored on the appropriate project hard drive.

4. Data Access

Access to actual data submissions/raw data is limited to researchers on the project whose names appear on the USC CSSE data access list. Researchers will sign nondisclosure agreements for Affiliates when requested by the Affiliate. The raw data will not be made available to Affiliates. Summary data will be made available only if data from at least four Affiliates is included in a

[2] Source/project identifying information excluded and only identification number described above used to index/identify data set.

data summary category. In addition, any summary data will not be dominated by a single source (e.g., over 50% of the data from one source). For those cases where there is a dominate source, data records will be randomly selected from the dominant source so that the summary criteria are met. The one exception to these rules is that the COCOMO™ project will provide each Affiliate with summaries of their own data.

For sanitized data records that have reached their "sunset time", records are incorporated in an online repository. Summary data is made available through this online repository in response to user queries that result in data from at least four data sources. This data is made accessible via username/password protections so that access can be controlled and monitored.

5. Getting Help

5.1 Phone Help Resources

5.1.1 USC Affiliate Questions

Direct questions on affiliate matters, meeting dates and locations, copies of proceedings, and general administrative matters to these people:

Julie Sanchez (jasanche@usc.edu)….....…….(213) 740-5703
Center for Software Engineering FAX…….(213) 740-4927

5.1.2 Questions About Data

For general questions or questions on the USC COCOMO™ model, data definitions, or project data collection and management, contact:

Barry Boehm…….. (213) 740-8163
Jo Ann Lane……... (858) 945-0099
Center for Software Engineering FAX……... (213) 740-4927
Internet Electronic-Mail……cocomo-info@sunset.usc.edu

For questions on the DACS-USC Repository, data definitions, or project data collection and management, contact:

Jo Ann Lane…….……... (858) 945-0099

USC CSSE Controlled Data Access List

Personnel currently on the access list include:

1. Dr. Barry Boehm
2. Dr. Jo Ann Lane
3. Dr. Brad Clark
4. Thomas Tan
5. Qi Li
6. Ramin Moazeni